DOCUMENT RESUME

ED 261 098 TM 850 525

AUTHOR Cashin, William E.
TITLE Student Ratings: The Need for Comparative Data.
PUB DATE 31 Mar 85
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Faculty; *College Students; *Comparative Analysis; Higher Education; Rating Scales; *Student Evaluation of Teacher Performance; Teacher Evaluation; Teacher Improvement; Test Reliability; Test Use; Test Validity
IDENTIFIERS Evaluation Problems

ABSTRACT

There are four reasons why comparative data is needed for student ratings of faculty performance: (1) the considerable inflation of student ratings; (2) the great variability in the way students rate different items; (3) because student rating systems must be flexible and comparable; and (4) because of factors (such as student motivation, class size and academic field) which bias student ratings. Studies involving the Instructional Development and Effective Assessment System, the Student Instructional Report, the Purdue University Cafeteria System, the Instructor and Course Evaluation System, and the Course/Instructor Evaluation Questionnaire are used to illustrate these problems with student rating systems. It is concluded that without comparative data, use of student ratings for teaching improvements is misleading and use for personnel decisions is inaccurate at best. A 24-item bibliography is included. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED261098

STUDENT RATINGS: THE NEED FOR COMPARATIVE DATA

William E. Cashin
Center for Faculty Evaluation & Development
Kansas State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Cashin, William E.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 525

Presentation made at Division J symposium on "Perspectives on Student Evaluation of Instruction: Lessons from Established Ratings Systems," American Educational Research Association Annual Meeting, Chicago, IL, March 31, 1985.

There is disagreement, at least among faculty, whether student ratings require comparative data. Some faculty argue that student ratings should be treated as criterion-referenced measures. So, for example, a rating of "3" on a 5-point scale is like a "C" grade and should be considered acceptable performance. Others suggest that, without comparative data, student ratings cannot be meaningfully interpreted. The term "comparative data" is being used rather than "norms" because no claim is being made that any of the data bases which will be referred to are representative of North America higher education in general.

I am going to argue for the need for comparative data for four reasons: first, because there is considerable inflation of student ratings; second, because there is great variability in the way students rate different items; third, because student rating systems must be both flexible and comparable, especially when the data are used as part of the basis for making personnel decisions; and fourth and finally, because there are a number of factors which bias student ratings. That is, there are a number of factors which "influence" student ratings other than the teaching effectiveness of the instructor.

RATING INFLATION

As I said, some faculty argue that 3.0 on a 5-point scale should be considered acceptable. The IDEA (Instructional Development & Effectiveness Assessment) system (Cashin & Perrin, 1978), the student ratings I am involved with at Kansas State, uses a 5-point scale. Using data from over 23,000 classes, the mean rating for the 38 IDEA items is 3.65. Since the average standard deviation is .605, this is about one standard deviation above 3.0. SIR (Student Instructional Report), available for the Educational Testing Service (ETS, 1979), is probably the most widely used student rating form in the United States. Using data from over 8,000 classes, looking at their 8 items which also use a 5-point scale, the average is 3.63, again about one standard deviation above 3.0. Similar inflation would be found on other student rating forms. What this suggests is that on most 5-point student rating items, a 3.0 cut off is not around the 50th percentile, but around the 16th. I think everyone using student rating data, whether for teaching improvement or for personnel decisions, needs to have that kind of information. I saw data for one student rating form developed by a fine arts department where the mean on a 5-point item dealing with the overall effectiveness of the instructor was 4.42--that was real data.

ITEM VARIABILITY

Although students tend to rate most items high on the average, there is still considerable variability in the way they rate individual items. Looking at the 20 teaching methods items on IDEA (Cashin & Perrin, 1978), there is 1.05 points difference between the means of the highest rated item and the lowest, a difference of 1.7 standard deviations. The mean of the highest rated item--the instructor "seemed enthusiastic about the subject matter"--is 4.3. The mean on the lowest item--the instructor "explained the reasons for criticisms of students' academic performance"--is 3.3.

On the twenty SIR teaching items (ETS, 1979), which use a 4-point scale, the difference between the highest and lowest item is .44, 1.5 standard deviation apart. The mean on the highest item--the instructor was well prepared for each class--is 3.4, on the lowest--the instructor

made helpful comments on papers and exams--3.0. Similar variation is found in other "well developed" student rating forms. By "well developed" I mean somebody with knowledge of educational measurement was significantly involved in the development of the form. The University of Washington system (Gillmore, 1980) now has seven different forms available to meet a variety of instructional approaches. On Form A (Gillmore, 1975), for small lecture-discussion type courses, the difference between the mean of the highest and lowest items is .71 (on a 6-point scale), 1.2 standard deviations. The highest mean--student confidence in instructor's knowledge--is 4.0, the lowest item--dealing with evaluation and grading techniques--is 3.3.

Without comparative data it is impossible to meaningfully interpret student rating data. Even if a large part of the variation in student ratings is due to actual differences in teaching effectiveness in the United States, for example, the three lowest items above all concern the degree to which we give students constructive feedback about their performance, it still makes a big difference if an instructor was highly rated on an item most everyone else does poorly on versus one where most teachers receive high ratings.

FLEXIBLE AND COMPARABLE

A major problem with many student rating forms is that they use a single set of items for all classes. Implicit in such forms is the assumption that there is a single, correct way to teach that should be used by all instructors in every class they teach. So, instructors should always speak clearly, be well organized, use discussion, and love the students. Personally, I believe that there are several effective ways to teach, and that student rating systems must accommodate those differences. There have been at least three different solutions to this problem. Perhaps the earliest was Purdue University's Cafeteria system (Purdue, 1974), which uses five general or summary items:

My instructor motivates me to do my best work.
My instructor explains difficult material clearly.
Course assignments are interesting and stimulating.
Overall, this course is among the best I have ever taken.
Overall, this instructor is among the best teachers I have known.

It then permits the instructor to choose up to 35-additional items (from a pool of 200) which the instructor considers appropriate for his or her class.

ICES (Instructor and Course Evaluation System) from the University of Illinois (Illinois, 1977), a more recent and more refined cafeteria system, uses three common global items:

1. Rate the course content
2. Rate the instructor
3. Rate the course in general

and lets the instructor select 23 additional items from the pool of over 400 items. Both Purdue's and ICES' cafeteria systems provide comparative data, Purdue for all items, ICES for the global and some common departmental items.

I think that the logic of cafeteria systems has much to recommend it. Flexibility is provided for improving teaching by the items the instructor chooses; comparability for personnel decisions is supplied by the few common global or summary items.

I suspect that this approach is also efficient. There have been factor analytic studies which have included global items. Issacson et al (1964) found four factors analyzing four different data pools. Their first factor, they called General Teaching Skills. Their first three highly loaded items were:

1. How would you rate your instructor in general (all-around) teaching ability? (Factor loadings ranged from .82-.86)
2. He put his material across in an interesting way. (.81-.88)
3. He stimulated the intellectual curiosity of his students. (.74-.85)

Their overall course item, however, "How would you rate the over-all value of this course?" did not load as highly (.34-.67).

Hartley and Hogan (1972) found seven factors. Their first factor they called Overall Evaluation which consisted almost entirely of global items:

Overall rating of course (.77)
Recommend it to friend (.70)
Overall rating of teaching (.68)
Would you take another course (.60)
Interesting presentation (.54)

A factor analysis of SIR (Centra, 1973) is less supportive of my argument. Centra found six factors. "Overall effectiveness of instructor" was the third item to load (.50) on the Factor III, Lectures. "Overall value of course to student" was the third item to load (.50) on Factor IV, Reading Assignments.

I consider the above studies suggestive at best, but my hypothesis would be that three to five global items would pick up from 50 to 70 percent of the variance that 20 to 40 specific student ratings items would account for in any general outcome or criterion measure. I am not sugesting that these global items will pick up all of the variance associated with specific teaching methods items, for example, certain discussion techniques, but neither will such discussion items correlate highly with outcome measures if a wide variety of college courses are sampled. So, if you collect ratings from multiple courses for multiple terms as is recommended (Gillmore, Kane, & Naccarato, 1978), you will obtain most of the useful information you need to make personnel decisions. Further, you will not be confused by data best interpreted by the individual instructor, faulting some instructors for not using discussion when they are primarily teaching facts, and faulting other instructors for not having organized classes when they are using discussions to develop critical thinking, change attitudes, or the like.

A second way to solve the flexibility-comparability problem is the approach the University of Washington (Gillmore, 1980) takes, using different forms for different types of classes.

IDEA (Hoyt & Cashin, 1977) solves the flexibility problem in a different way using students' reports of their progress on course objectives as the primary measure of teaching effectiveness. But, the importance of each objective is weighted by the instructor or department, so different criteria are used for different classes.

FACTORS WHICH BIAS

The question here is whether there are variables which significantly influence student ratings and which are not primarily reflective

of the instructor's teaching effectiveness. The good news is that in general the following variables show little or no relationship to student ratings (Menges, 1973; Feldman, 1977; McKeachie, 1979):

age of student
sex of student
sex of instructor
students' level (freshman, etc.)
students' GPA
time of day when course is given (Alemoni, 1981).

The bad news is that there are other variables which do appear to bias ratings. Some of these problems, like ratings being higher if the instructor remains in the room while the students are completing the ratings, are easily solved by not having the instructor remain in the room.

I would like to address three variables: student motivation, size of class, and academic field. These are perhaps the most "biasing" variables (or ones which at least show higher correlations with student ratings), but which can be controlled by using comparative data.

The first is student motivation (desire to take the course, or elective vs. required course). More motivated students (or elective courses) receive higher student ratings (Costin, Greenough, & Menges, 1971; Feldman, 1978; Kulik & Kulik, 1974; Kulik & McKeachie, 1975; Marsh, 1984; McKeachie, 1979).

IDEA (Hoyt & Cashin, 1977) uses one item to control for student motivation: "I had a strong desire to take this course." That one item picks up from 10% to as much as 42% of the variance in most other IDEA items (Cashin & Perrin, 1978), for example:

Introduced stimulating ideas about the subject (30%)
Discovering the implications of the course materials for understanding myself (interests, talents, values, etc.) (27%)
As a result of taking this course, I have more positive feelings toward this field of study (42%)

The CIEQ (Course/Instructor Evaluation Questionnaire) developed by Alemoni and available for the University of Arizona provides three norm tables for required vs. elective courses (Arizona, 1976): courses required for 0% to 33% of the class, 34% to 66%, and 67% to 100%. Comparing the scores on the five subscales from the two extreme tables, the average difference is .5. These are items using a 4-point scale.

A relationship has also been found between the size of the class and student ratings. Smaller classes tend to receive higher ratings (Kulik & Kulik, 1974; Kulik & McKeachie, 1975; Feldman, 1978, 1984; McKeachie, 1979). IDEA controls for class size by using comparative data for four class sizes, ranging from less than 15 students to 100 or more students. IDEA also controls for student motivation at the same time. We reason that it is easier to teach small groups of students who want to take your class than it is large groups of students who do not, and that fact should be taken into consideration when interpreting your student ratings.

One variable that is not directly controlled by any student rating system is academic field. There is growing evidence that humanities and fine arts courses receive higher ratings than social science courses which in turn receive higher ratings than math and science courses

(Kulik & Kulik, 1974; Kulik & McKeachie, 1975; Feldman, 1978; McKeachie, 1979). Presently SIR provides the most comprehensive comparative data with tables for 33 fields for four-year colleges (ETS, 1979) and tables for 31 fields for two-year colleges (ETS, 1977).

IDEA has unpublished means for 25 academic fields. On a widely chosen course objective like gaining factual knowledge the means only vary .4, but on objectives like general liberal education, they vary as much as 1.7 tenths of a point, that's two and a half standard deviations.

CONCLUSION

The point of everything I have been saying is that student ratings vary, often widely, in relationship to a number of variables. Comparative data permits you to get some kind of a handle on these differences. I am not suggesting that simply providing comparative data will lead to significant improvement of teaching, indeed, the research suggests that providing some kind of consultation along with student rating feedback is far more productive (Cohen, 1980). To try to use student ratings for teaching improvements without comparative data, however, is misleading at best and at worst is meaningless. To use absolute ratings for personnel decisions is inaccurate at best, and at worst is both unethical and probably illegal.

References

Alemoni, L.M. (1981). Student ratings of instruction. In J. Millman (Ed.), Handbook of teacher evaluation. Beverly Hills: Sage.

Arizona, University of. (1976). Norms based on proportion of students taking the course as a required vs. elective. Tucson: Author.

Cashin, W.E., & Perrin, B.M. (1978). IDEA technical report no. 4: Description of IDEA Standard Form data base. Manhattan: Kansas State University. Center for Faculty Evaluation and Development.

Centra, J.A. (1973). The Student Instructional Report: Comparisons with alumni ratings, item reliabilities, the factor structure. (No. 3). Princeton, NJ: Educational Testing Service.

Cohen, P.A. (1980). Effectiveness of student rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education, 13, 321-341.

Costin, F., Greenough, W.T., & Menges, R.J. (1971). Student ratings of college teaching: Reliabilty, validity, and usefulness. Review of Educational Research, 8, 511-535.

Educational Testing Service. (1977). Student Instructional Report: Comparative data guide for two-year colleges and technical institutions. Princeton, NJ: Author.

Educational Testing Service. (1979). Student Instructional Report: Comparative data guide for four-year colleges and universities. Princeton, NJ: Author.

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education, 6, 233-274.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. Research in Higher Education, 9, 199-242.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. Research in Higher Education, 21, 45-120

Gillmore, G.M. (1975). Statistical analysis of the data from the first year of use of the student rating forms of the University of Washington Instructional Assessment System. Seattle: University of Washington, Educational Assessment Center.

Gillmore, G.M. (1980). A brief description of the student rating forms of the University of Washington Instructional Assessment System. Seattle: University of Washington, Educational Assessment System.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. Journal of Educational Measurement, 15, 1-13.

Hartley, E.L., & Hogan, T.P. (1972). Some additional factors in student evaluation of courses. American Educational Research Journal, 9, 241-250.

Hoyt, D.P., & Cashin, W.E. (1977). IDEA technical report no. 1: Development of the IDEA system. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Isaacson, R.L., McKeachie, W.J., Milholland, J.E., Ling, Y.G., Hofeller, M., Bairwaldt, J.W., & Zinn, K.L. (1964). Dimensions of student evaluations of teaching. Journal of Educational Psychology, 55, 344-351.

Illinois, University of. (1977). ICES item catalog. Urbana-Champaign: Author.

Kulik, J.A., & Kulik, C.C. (1974). Student ratings of instruction. Teaching of Psychology, 1, 51-57.

Kulik, J.A., & McKeachie, W.J. (1975). The evaluation of teachers in higher education. In F.N. Kerlinger (Ed.), Review of research in education (Vol. 3), Itasca, IL: F.E. Peacock.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

McKeachie, W.J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384-397.

Menges, R.J. (1973). The new reporters: Students rate instruction. In C.R. Pace (Ed.). Evaluating learning and teaching: New directions for higher education. No. 4. San Francisco: Jossey-Bass.

Purdue Research Foundation (1974). Instructor and course appraisal: Cafeteria System. West Lafayette, IN: Author.